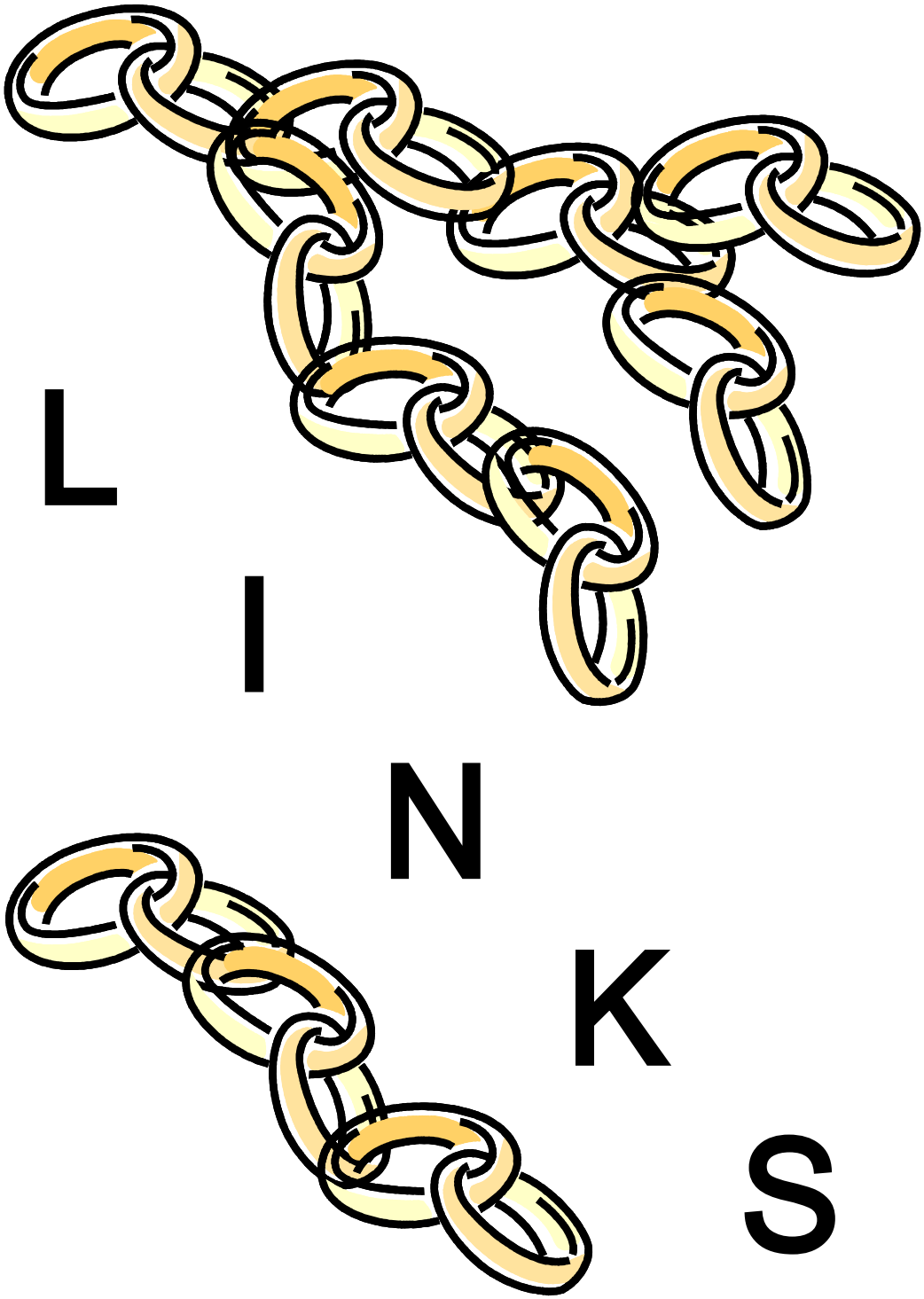


LINKing System for historical family reconstruction

1a **Link**ing **s**ystem for historical family reconstruction

1b **LINKS**

1c **Prof. dr Kees Mandemakers**

International Institute of Social History
Cruquiusweg 31
1019 AT Amsterdam
020-6685866
kma@iisg.nl

2 Summary

LINKS aims at reconstructing all nineteenth and early twentieth century families in the Netherlands. This reconstruction will be based on GENLIAS, which is a digitized index of all civil certificates from this period. For fifteen years numerous volunteers have been working to build the index, which contains not only the names of born, deceased and married persons, but also the names of their parents, places of birth, ages and partly their occupational titles.

The availability of this dataset offers an enormous potential for scientific research provided that individuals are linked into families. This is not only of the utmost importance for historical demography and social and economic history but also for onomastics, epidemiology, anthropology, historical sociology and genetics.

As a consequence of the high degree of fuzziness of the spelling of both first and last names (because of errors, mistaken statements and inconsistencies during registration, regional deviations, indexation, etc.), and inconsistencies between archives in local data storage, linking is a complicated task. LINKS has formulated three requirements for successful reconstruction and dissemination: a) a dynamic parser which converts the input from GENLIAS into a standardized data structure, b) nominal record linkage procedures with self learning capacities and c) a retrieval system including GIS-references and visualization procedures. Because the GENLIAS database is continuously renewing and extending its content, there will be new LINKS output releases annually. Not only for research but also for the GENLIAS community which in return will receive families built by the LINKS software.

3 Classification

I **Interoperability of large scale and distributed sources**

4 Composition research team

Prof. dr K. Mandemakers	Senior researcher Professor	Large historical databases	International Institute of Social History (IISH), head Historical Sample Netherlands (HSN), Amsterdam Erasmus University Rotterdam (Faculty of History and Arts)
Drs. F.P. Bosmans	Head informatics	Administrator central database GENLIAS	Tresoar (Fries historisch en letterkundig centrum), Leeuwarden,
Dr. G. Bloothooft	Senior researcher Assistant professor	Nominal record linkage	Meertens Instituut, Amsterdam University Utrecht (Utrecht Institute of Linguistics, OTS)
Dr. H.J. Hoogeboom	Assistant professor	Natural Computing	Leiden University (Institute of Advanced Computer Science, LIACS)
Dr. D.P. Huijsmans	Assistant professor	Spatio-temporal exploration	Leiden University (Institute of Advanced Computer Science, LIACS)
Dr. J. Kok	Senior researcher	Family History	Virtual Knowledge Studio (VKS), Amsterdam
Prof. dr J.N. Kok	Professor Supervisor PhD	Algorithms and Foundations of Software Technology	Leiden University (Institute of Advanced Computer Science, LIACS),
Dr. W.A. Kosters	Assistant professor	Artificial Intelligence; Data Mining	Leiden University (Institute of Advanced Computer Science, LIACS)
Drs. P. den Otter	Head informatics	Secretary GENLIAS assembly	Historisch Centrum Overijssel (HCO), Zwolle.
Dr. F.W.A. van Poppel	Senior researcher	Historical Demography	Netherlands Interdisciplinary Demographic Institute (NIDI), Den Haag
Vacancy	PhD		Leiden University (Institute of Advanced Computer Science, LIACS)
Vacancy	Postdoc		To be divided in three positions, see section 6b
Vacancy	Programmer		International Institute of Social History (IISH), Amsterdam

5 Research school

Institute for Programming Research and Algorithmics (IPA).
Netherlands Graduate School of Linguistics (LOT).
The N.W. Posthumus Institute (Research School for Economic and Social history).

6 Description of the proposed research

GENLIAS

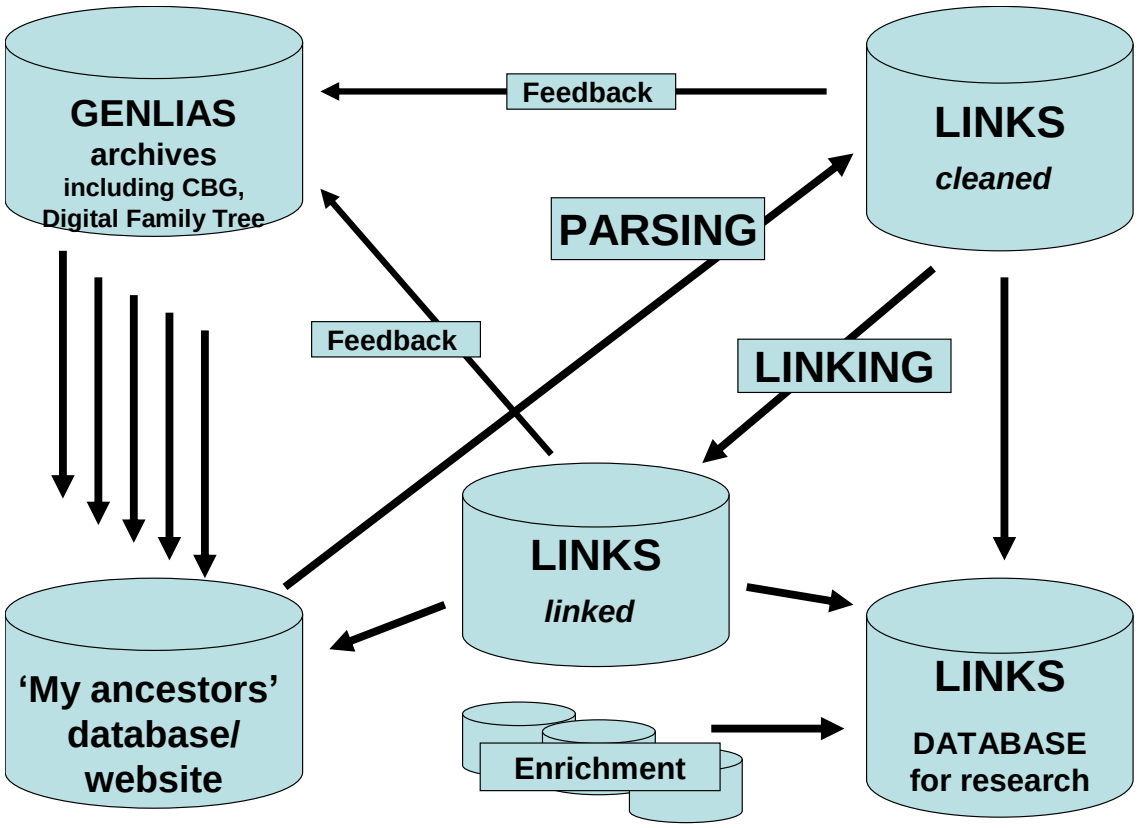
The system of vital registration in the Netherlands is a legacy of the Napoleonic period. It prescribes and regulates registration of births, marriages and deaths, and has resulted in a huge and invaluable source of information for scholars and the general public. In the mid nineties a large-scale digitization initiative, called GENLIAS (www.genlias.nl), started as a cooperation between twelve regional archives under national coordination. GENLIAS, which is being built with the help of numerous volunteers, currently holds a database with key information on certificates of marriage, death and birth from 1811 until the (shifting) year that these data become publicly accessible. This time limit, 100 years for birth certificates, 75 years for marriage certificates and 50 years for death certificates is regulated by the Law covering Dutch vital registration. GENLIAS will eventually contain about 32 million certificates. Currently, it contains key information from about 11 million certificates with 45 million entries of individuals. Central focus is on marriage certificates which now covers 85% of all potential certificates and is expected to be complete in 2010. The regional archives of Zeeland, Friesland and Drenthe have almost finished data entry work of all certificates. Given the present speed of data entry and the importance attributed to GENLIAS by the archives, we can expect that within the coming 15 years all potential certificates will be entered.

The GENLIAS database is one of the largest in the world with respect to digitizing handwritten historical sources. Last year, the GENLIAS website welcomed four million visitors searching for genealogical information. During 2009-2013 a national platform for genealogical research will be developed. The project is called 'Mijn Voorouders' (‘my ancestors’), and will merge and extend the functionality and content of the three main Dutch websites with genealogical information. The project is based on a budget of m€ 3,9.[1] The new online platform will entail possibilities for distributing more information than the index itself, for example links to other certificates or pedigrees and family reconstructions.

In the LINKS project we will build an interoperable system for delivering GENLIAS data to the scientific community in an enriched and linked way. We will give feedback to the GENLIAS community to improve the database and deliver the links to build pedigrees and families. In figure 1 an outline of this process is sketched which will be elaborated upon below. Although the LINKS system will be built for the projected situation in 2012 including all marriage certificates and about half of the birth and death certificates, the software will be designed with the prospect of including all publicly available certificates in 2022.

[1] The new website will be a combination of the now existing GENLIAS.nl, deDigitaleStamboom.nl and CBG.nl. About 2 m€ will be contributed by the archives while the remainder, m€ 1,9, is asked from the PRIMA fund of the Ministry of Education, Culture and Science. The Ministry decided positively on a first part of € 367,000 for the first project year and intends to continue subsidizing.

Figure 1 Outline of the data flow between GENLIAS and LINKS: parsing, linking, giving feedback and distributing data from the Dutch civil certificates.



6a Scientific aspects

Parsing GENLIAS

GENLIAS has been designed as a decentralized system in which archives agreed to deliver data to the central database: at least all first and last names from key persons in the certificates and source identifying information including date and municipality. The archives could make their own decisions as to what other data are entered into the index and in what way. Quite often a free format field is used to enter age or date of birth, civil status, or former spouses. Occupational titles are only systematically entered by the archives of the provinces of Groningen, Overijssel, Gelderland, Zeeland and Limburg. Although in the future we can expect a more uniform data model, within the LINKS project we have to build a parser suited for several different data models at the input side.

The parser has to interpret all data, both standard and non-standard, from GENLIAS and will convert the data into a standard LINKS format for further processing. The parser must first convert the standard data fields, performing simple tasks like unifying the representation for accented characters, while flagging unknown fields (which are currently marked in different ways). Secondly, the parser must take care of the free format field(s). This approach needs (mild) natural language processing techniques. It is clearly not possible to fully exploit the semantic contents, but especially in the limited

domain of GENLIAS records there are several near standard constructs. For example, information on former weddings, the age of the partners or their occupations is often included as comments. But a free format field may also contain observations made by the volunteer on invalid, contradictory or unreadable information. This pattern detection task nicely fits with the data mining activities of the later stages.

The output of the parser will lead to feedback to the GENLIAS partners. This may involve general suggestions to improve the uniformity of the data, but also specific instructions to verify the data in certain certificates. As a matter of fact, unreadable certificates may be checked against the second copy that is archived elsewhere.

Nominal record linkage

The basic problem that has to be solved is the identification of individuals and their family relations on the basis of the civil records. This process, referred to as nominal record linkage, is complicated because names cannot be used as unique identifiers for persons. One and the same person may occur with different names, and a single name may refer to many persons. An extensive literature on this subject is available, see, e.g. Goiser and Christen 2006, Herzog et al 2007, Schürer 2007 and the references therein.

The basis for our linking process will be the *combination of the names of couples*: bride and groom, mother and father. Because in the Dutch civil administration everyone (also women) keeps his or her first name and surname given by birth, the combination of names of married couples are quite unique, especially if linking can be limited within ranges of time bounded by years of birth, periods of reproduction and life span. The combination of the full names of a couple forms the spine of our family reconstruction, as the parents need to be mentioned in records of birth, marriage and death. We link a) marriages to marriages (bride and groom to the marriage of their respective parents), forming pedigrees, b) births to marriages (of parents) forming families and c) deaths to marriages (with spouse(s) and from parents) forming small life histories. In general we will work by linking couples of two (married) persons. Linking birth certificates with marriage and death certificates will even use combinations of three to four names: own name, name of spouse and names of the parents.

Common issues are the handling of spelling errors, errors in dates and ages, typing errors during digitization, and missing information due to incomplete digitization of information present in the original records (Bloothooft, 1998). The record linkage procedure will be tuned in such a way that it will start with accepting linkages of couples which are fully identical under realistic constraints of periods of reproduction. This means that ages of mothers will be compared to ages of brides and grooms, birth dates of children etc. As a next step fuzzy (approximate) matching will be applied in three ways: a) using algorithms to overcome spelling variation, b) using self learning techniques and c) making use of regionally based differences in name giving, spelling etc.

A) *Spelling variations*. We will use string comparison and pattern matching techniques. Most well known is the Levenshtein edit-distance, which is a well-defined measure of difference between strings. It efficiently computes the minimum number of insertions and deletions necessary to transfer one object into another. It allows many adjustments, e.g. for weighting special combinations of characters or assigning costs that depend on the probability of occurrence. It is often used in linguistic context, but also in bioinformatics, e.g. for DNA. Several measures are proposed and discussed (Laros and Kosters 2007). Also hybrid approaches exist that are based on a combination of phonetic and string matching techniques. It turns out that the choice of such a measure is a complicated task, suited for an expert in the data domain and a computer scientist. See Christen (2006) for an extensive list of possible approaches and a discussion of some preliminary experiments.B) *(Un)Supervised learning techniques*

(Russell and Norvig 2003). The Levenshtein algorithm does not include knowledge about the domain, like similar sounds that are written in various ways. This knowledge can be added to the algorithm by introducing weights. In the past this was done by compiling tables of substitutions ("ei" = "ij", "c" = "k"). Similarly, typing mistakes can be valued by giving mistakes with neighboring keys on a keyboard less weight than more random ones. One has also to accommodate the existence of highly differing (informal) first name variants such as Jan and Johannes, or Guillaume as a French translation of Willem. Again, by giving small weights to these kinds of variation we can improve the linkage system. We can work with lists of these similarities, but we can also learn from the database how often links are blocked by this kind of differences. We will include techniques that are tailored to automatically learn which kind of spelling variations is quite common. For instance, counting what kind of combinations of first names and surnames (of couples) occur in the certificates is the basis for such a process. More in general, we will explore the possibilities of data mining techniques (Tan et al., 2006) to extract knowledge from the database. Some of these can be applied to improve linking by discovering unexpected dependencies in the information present in the database.

C) Regional variation. It is expected that there are considerable regional differences in socio-demographic and onomastic issues, such as the use of patronymics, the introduction of surnames after 1811, naming traditions, family relations, and migration distances that should be considered in the linkage process. Also, from the administrative point of view, in the period before 1880 practices proved not as standardized as supposed by the national state, and regional variation in the way of registration can be anticipated. This is especially likely in regions like Limburg with a lot of foreign influence. The first two approaches can be combined with techniques using spatial and temporal components of the data, which allows for the usage of specially designed rules that apply in certain provinces or within certain distances and may change over time. This can be either user-originated or inspired by computational methods.

Pruning. The comparison of all couples on all 32 million certificates in order to evaluate them for possible matches is very time consuming and in practice unachievable. This is a common problem in record linking. Measures must be taken to have a system that does not need to compare all records, and is able to decide quickly for the majority of the records that they cannot match. This is also referred to as blocking. We will combine and test various techniques. We will only match within logical time ranges and start e.g. with matching identical couples (four identical names: two first names and two surnames), followed by matching all instances where any three of four names in a couple are identical, while evaluating the fourth name against minimal distance measures. To minimize computer time we will combine this approach with pre-matching all occurring surnames and first names, since the number of unique names is considerably smaller than the total number of names. Once these easy matches are set aside, the remaining (problematic) records can be added to the existing structure. Experiments will have to prove the usefulness of the linking techniques mentioned before and the order in which they can be applied best.

*Statistics***.** The aim of the linkage process is to add a degree of certainty to each link. This certainty is based on likelihood of spelling variation, but also on likelihood of the life cycle, and to some extent geographic distance. In the time domain, we need estimates for age of marriage, (conditional) duration of life, and so on. In most cases, these statistics can be derived directly from the data. Geographical distance is known beforehand (but may become complicated if expressed in travel time). The weighting of the evidence from the different domains of spelling, time and geography into a single measure of certainty has to be investigated in an empirical way. In all cases, our procedure will result in an ordered series of hypotheses from high to lower likelihood. This can help genealogists to guide efforts to find additional evidence. From a scientific point, such an effort is usually not feasible, but the best matches themselves may already provide a solid basis for further large scale analyses.

Evaluation. The quality of the resulting family reconstruction can be tested against the 78.000 persons from the birth period 1812-1922, constituting the Historical Sample of the Netherlands (HSN). The HSN is based on birth certificates and enriched with marriages, deaths and information from the population registers (Mandemakers, 2000). This benchmark with already established links can be used to estimate the amount of under- and overmatching, and thus identify weaknesses in the automatic linkage process and will be used to suggest approaches to counter these problems.

Feedback. The whole process of linking the GENLIAS data must be seen as an iterative process that will give better results with every new data release from GENLIAS and with evaluating every new rule. Part of the errors found will have originated from reading or typing errors during the data entry in the archives. These errors will be listed and communicated to the GENLIAS community, and they can be used to improve the content of GENLIAS

Exploring the database

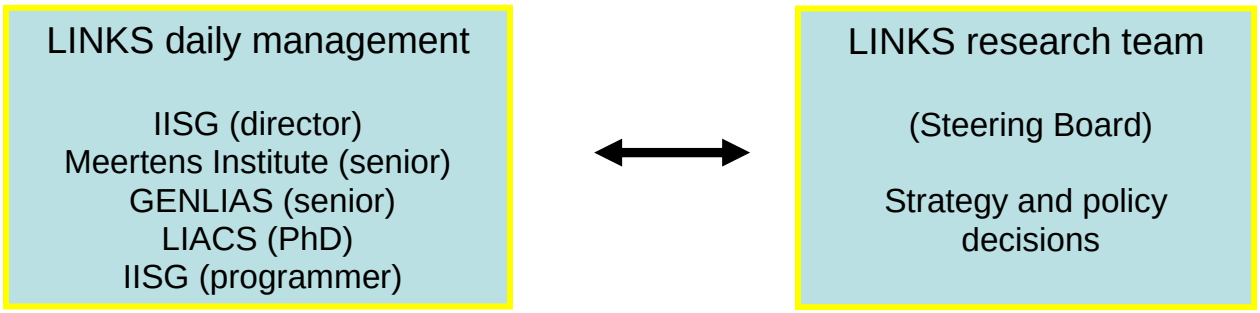
The third part of the research is the development of an interactive environment for researchers, built upon efficient indexing, geo-referencing and visualization. Even when fully linked, with 100% reliability, the database will be of little use if it cannot be searched efficiently. The scientific user must be able to efficiently perform SQL-type queries to the data in order to extract the required information. Searching is not the main concern of this project, but is worthwhile to investigate when searching problems can be solved by applying the techniques that accommodate the linking phase. Visualization tools can be helpful, even in the production phase, to access the large amount of data. In later stages researchers can benefit from software to exploit the spatio-temporal information in the data.

We will develop an environment in which a researcher can make his/her own selections and projections from the database. The database will be formed by combining the linking information with the original dataset cleaned by LINKS. Standard enrichment data sets such as GIS-codes, contextual codes on municipalities and HISCO-codes for occupational titles will be integrated as well (see figure 1). By explicitly geo-referencing data records, learning and knowledge representations can be made to encompass temporal and geographical components. Exploration and data mining will be done in a graphical interface allowing easy manipulation based upon spatio-temporal criteria. This will make it possible for example to display on the screen regions with high frequencies of cousin-marriages or the spread of certain variations in names during a specific period. Application of these criteria to large historical databases has been quite limited so far and will have an innovative character. Especially the mining of spatio-temporal patterns, the development of application specific ontologies and the extension of markup languages to spatio-temporal ones form an active research area that we would like to contribute to during this project (Huijsmans and Smeulders 1999; Koubarakis and Sellis 2003).

6b Multidisciplinary cooperation

Figure 2 gives an overview of the management structure and the way the cooperation between the different disciplines is organized. The work is directed by the IISG. Major decisions concerning the direction of the project will be taken by the senior members of the research team (see section 4) which will function as a steering board. The research team forms a liaison between the cultural heritage part of the project, the scientific community and the informatics part. The group will meet at least four times a year.

Figure 2. Management structure LINKS.



Research team

The team is based in seven institutes: the International Institute of Social History (IISG), Historisch Centrum Overijssel (HCO), Tresoar, the Leiden Institute of Advanced Computer Science (LIACS), the Netherlands Interdisciplinary Demographic Institute (NIDI), the Meertens Institute and the Virtual Knowledge Studio (VKS). LIACS, NIDI and VKS can be considered as research institutes, HCO and Tresoar as cultural heritage and IISG and Meertens cover both sectors and are internationally important institutes in the field of humanities. IISG, Meertens Institute, NIDI and VKS are part of the Royal Netherlands Academy of Arts and Sciences (KNAW).

The Historisch Centrum Overijssel hosts the executive secretariat of the GENLIAS assembly, representing all archives involved in GENLIAS. Tresoar organizes the data flow from all databases into the central database of GENLIAS.

The Meertens Institute is specialized in Dutch family and first names and will play a vital role in validating the record linkage process.

The GENLIAS-data are already available for scientific research and are distributed by the Historical Sample of the Netherlands (HSN) which is located at the International Institute of Social History (IISG). The IISG has acquired from the GENLIAS archives the distribution rights for scientific research and distributes the data by way of a license system (see section 6c for first research results). The IISG will direct the LINKS-program and will host the programmer and the LINKS-database. Because of the decentralized character of GENLIAS the IISG in combination with the HCO, Tresoar and the Meertens Institute can be considered as the cultural heritage partner in this application.

LIACS (Leiden University) is represented by two research groups. The ‘Algorithms group’ works in the field of data mining and incorporates methods from Artificial Intelligence in practical problems, especially when large data sets are involved. This research often relies upon methods that originate in nature, such as neural networks, evolutionary computing and computational biology (inspired by matching and querying huge data sets to localize patterns). The ‘Imagery and Multimedia

group' aims at maximizing the role that time and location data in large historic databases can play for researchers who perform data mining, visualization and record linking.

The Netherlands Interdisciplinary Demographic Institute conducts scientific research into population issues and disseminates demographic knowledge and information among the scientific community, policymakers and society at large. Their representative dr. F. van Poppel studied 19th century marriage patterns and is a famous specialist on historical demography.

The Virtual Knowledge Studio supports e-research, among others by helping to design web interfaces for interdisciplinary use. Their representative dr. J. Kok is a well-known specialist on family history.

Management

Given the decentralized character of GENLIAS and the scientific community the project will be managed by a team of three persons based at IISG (prof. K. Mandemakers), the Meertens Institute (dr. G. Bloothooft) and by a senior from one of the main GENLIAS archives. It was considered too difficult to find one person able to combine both coordination and dissemination of the results in disciplines as diverse as onomastics and historical demography. The IISG already hosts the GENLIAS-data for scientific distribution and more in general has experience in developing data for scientific research. By way of the network of the HSN it is equipped to distribute the data among other scientific fields as well. Because of the decentralized character of the GENLIAS database we need a person who knows all the archives and persons to be addressed in the process of giving feedback and receiving new editions of the data sets. The onomastics group of the Meertens Institute hosts standards of databases of first names and surnames, including knowledge about regional variations, while dr. G. Bloothooft is also a specialist in nominal record linkage.

The project is directed by the IISG. The IISG and her ICT department will employ the programmer. The PhD will be based at LIACS and will be supervised by prof. J.N. Kok. Candidate PhD is Maarten Oosten MSc who was already a trainee at IISG linking GENLIAS and who shows serious interest to continue his research. The programming will be directed by prof. K. Mandemakers in cooperation with the PhD and the ICT department at IISH. The PhD will also have a major role in the development of the functional and technical design and will do part of the programming. In case of differences of opinion about directions of programming the director and the supervisor PhD will discuss the case. If they cannot reach a solution the director will decide after consultating the Steering Board.

The management team will communicate with the archives and with the GENLIAS community by realizing common data formats and by publishing the results. They will lead the semantic aspects of the record linkage process whereby the PhD will concentrate on aspects of general informatics. The team will have combined meetings at least once a month. The team is also responsible for disseminating the data to the relevant research fields. This will done by publicizing and reporting at scientific congresses about the progress of the LINKS-project and by doing research with the database (co-) authoring scientific publications.

6c Relevance

Linking all certificates will ultimately allow for a family reconstruction of virtual all people born in the Netherlands between 1811 and 1922. The GENLIAS-community is now building a new web-interface for the general public. The results of LINKS will be used to add specific pedigrees and family reconstructions including likelihoods of alternative interpretations. This will give an enormous impetus to the work of countless local historians, genealogists, biographers et cetera.

LINKS will also strongly stimulate a number of research strands in the humanities. Firstly, the study of (regional variation in) kinship systems will be enhanced, as the project yields a sufficient number of three generation pedigrees to analyze cousin marriages, cross-sibling marriages etc. This is relevant for ethnologists as well as for population geneticists studying the effect of kin marriages on survival of children. Secondly, a variety of key issues in epidemiology and demography, for which information on the family context of mortality is indispensable (death clustering within families, grandmother hypothesis, effect of ages of mother at birth of first and last child on post-reproductive mortality) can be studied on a large scale (including stillborns who are included in the death certificates). Thirdly, adding location information to the certificates in GENLIAS offers new data for studies of (the family context of) internal migration and of the integration of marriage markets during the 19th century. Fourthly, the added information on occupational titles is very important for the study on social mobility and the question when and how The Netherlands have become a more 'open' society with equally distributed chances of social improvement. Also, it will now be possible to study the impact of wider kinship ties on individual mobility. Naming patterns can be studied as indicators of processes of integration, secularization and distinction.

The family reconstruction will generate valuable onomastic corpora, describing the inventory and frequency distribution of first names and surnames in the 19th century (including regional variations). It opens, for instance, the way to trace back the continuous process of renewal of the first name inventory (among which the use of more than one first name), and to study the beginning and spread of these mechanisms. This significantly extends onomastic research based on the current vital registration with more than a century. Also, traditions in naming after relatives can be easily studied on the basis of known family relations. An understanding and description of the way surnames were adopted in the early 19 century, which strongly varied regionally, is highly interesting from an onomastic point of view, but will also be incorporated in the family reconstruction procedures.

By combining with other datasets, GENLIAS' impact on the humanities research is further enlarged. Contextual information on municipalities will be provided by the Hub for Aggregated Social History (HASH). This database is now under construction at Radboud University (grant NWO Middelgroot). Linking the project's results with the Historical Sample of the Netherlands (HSN) will result in considerable extra information for the 78.000 HSN research persons, especially on marriages of relatives (parents, siblings, children). Already, the research group connected with the HSN has explored the potential of family reconstruction for the Eastern provinces and Zeeland which already reached 100% coverage of digitization and where the certificates include besides date and names also information on age, occupation and place of birth. Based on marriage certificates, changes in the occupations of woman were studied (Maas and Van Leeuwen, 2006) and the intergenerational transfer of age at marriage (Van Poppel et al, 2008a). Other studies, for example on mortality trends and changes in the geographical horizon of marriages, are under way (Van Poppel et al, 2008b and 2008c).

In the future other links with databases are anticipated such as Social Statistical Database which is based on the current Dutch population register, the central archive of causes of death (from 1937 onwards) and the archive of Personal Cards at the Central Bureau for Genealogy (CBG). Linking with these databases is allowed under strict conditions of privacy protection and will be of enormous importance for fields like epidemiology and genetics.

7 Description of the proposed plan of work

We will build a system that consists of the three main components as elaborated upon in section 6 (see also figure 1):

1 A parsing system to read the GENLIAS data and convert them into the data structure of LINKS.
2 A linking system to identify families and individuals in GENLIAS.
3 An output-system producing an enriched scientific database of GENLIAS resulting from the parsing and linking process.

The emphasis of the research of the PhD will be on the design and implementation of the linking system and aspects of data mining (indexing, querying and visualization). Questions to be solved here will lead to new insights in the field of computer science. We expect the researcher to report on implementation issues (in particular the choice of proper data structures that speed up the access time of the database), on fuzzy matching techniques (suitable for the domain of personal names used in GENLIAS) combining old techniques with new approaches, and on probabilistic rule sets (deciding between proposed links on logical and probabilistic grounds). The data will be XML-coded, we will explore EBXML which Dutch government is now implementing for her administration including the population register and the civil certificates.

We will work in an iterative way. That means that already during the first two years prototypes of all components will be presented to a wider audience. The following gives a detailed list of deliverables and time table. Main part of the deliverables includes documentation and the building and maintenance of a dedicated website. The deliverables will be the result of an intensive cooperation within the management team, e.g. the PhD will do most of the work of the functional and technical design, but the several versions will be extensively discussed within the management team before approval and other members of the team will also write parts of the design.

Component	Year	Month	Deliverable
1 (parser)	2009	5	First version of the documentation on systematic shortcomings in the GENLIAS data including algorithms to clean and disentangle data and to provide estimations for incomplete data (for example birth ranges on the basis of age and date of the certificate), where necessary specified per archive.
1	2009	5	First version of the data model for the cleaned GENLIAS database.
1	2009	9	First version of the parsing system, including documentation describing deliverables for archives (lists of possible mistakes, queries for systematic improvements).
2 (linkage)	2009	12	First version of the linkage system (baseline without fuzzy matching), based on marriage certificates.
2	2010	3	Second version of the linking system, concentrating on overcoming spelling variations by Levenshtein and introduction of self learning techniques, based on marriage certificates.
3 (output)	2010	6	First version of the output system (pedigrees and families), for GENLIAS and for scientific use, including keys for enriching the system.
2	2010	9	Third version of the linking system (now including knowledge from

			regional variations, based on knowledge Meertens Instituut, sophisticating the whole system of spelling variations and learning techniques), based on marriage certificates.
2	2010	12	Extension of the linking system to birth and death certificates (first version).
1-3	2011	1-12	Further developing and testing of in between versions of the three components.
1	2011	12	Final version of the parsing system (improved a.o. on the basis of feedback from the archives and the results from the linking system).
3	2012	3	The Meertens Institute will produce data sets with an overview of normalized family names and first names and how these names change over time and regions in the Netherlands.
2	2012	9	Final version of the linking system.
2	2012	9	Concluding conference for the GENLIAS community presenting the linkage information.
3	2012	12	Final version of the output-system, including XML coding and a website with tools for querying and visualization of results.
3	2012	12	Concluding conference for the scientific community presenting new studies based on the database.
1-3	2012	12	Thesis PhD (Book).

8 Literature

8a References

Bloothooft, G. (1998). Assessment of Systems for Nominal Retrieval and Historical Record Linkage. *Computers and the Humanities*, 32, 39-56.

Christen, P. (2006). A Comparison of Personal Name Matching: Techniques and Practical Issues. In *Proceedings of the Workshop on Mining Complex Data (MCD*) (Hong Kong: ICDM), 290-294.

Goiser, K. and P. Christen (2006). Towards Automated Record linkage. In *Proceedings Fifth Australasian Data Mining Conference* (Sydney), 61, 23-31.

Herzog, T.N., F.J. Scheuren and W.E. Winkler (2007). *Data Quality and Record Linkage Techniques* (New York: Springer).

Huijsmans, D.P. and A.W.M. Smeulders (eds.) (1999). *Visual Information and Information System. Proceedings ConferenceVisual 99* (New York: Springer).

Koubarakis, M. and T. Sellis (eds.) (2003). *Spatio-Temporal Databases, the Chorochronos Approach* (New York: Springer).

Laros, J.F.J and W.A. Kosters (2007). Metrics for Mining Multisets, In M. Bramer, F. Coenen, M. Petridis (eds.), *Research and Development in Intelligent Systems XXIV, Proceedings of AI-2007, the Twenty-seventh SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence* (New York: Springer), 293-303.

Maas, I. and Marco van Leeuwen (2006). Over dienstboden, landarbeidsters en andere werkende vrouwen. Beroepen van jonge vrouwen en hun moeders in de huwelijksakten van de Zeeuwse Burgerlijke Stand. *Zeeland* 15, 44-59.

Mandemakers. K. (2000). The Netherlands. Historical Sample of the Netherlands. In P. Kelly Hall, R. McCaa, G. Thorvaldsen (eds.). *Handbook of International Historical Microdata for Population Research* (Minneapolis: Minnesota Population Center), 149-177.

Poppel, F. van, C. Monden and K. Mandemakers (2008a). Marriage Timing over the Generations. *Human Nature* 19, 1, 7-22.

Poppel, F. van, H. van Dalen and E. Walhout (2008b). Diffusion of a Social Norm: Tracing the Emergence of the Housewife in the Netherlands, 1812-1922'. *Economic History Review*, in press.

Poppel, F. van, R. Jennissen and K. Mandemakers (2008c). Social Class and Adult Mortality: Indirect Estimation of Time Trends for the Netherlands (1820-1920). *Social Science History*, in press.

Russell, S. J. and P. Norvig (2003). *Artificial Intelligence, A Modern Approach* second edition (Upper Saddle River NJ: Pearson Education).

Schürer, K. (2007). Creating a Nationally Representative Individual and Household Sample for Great Britain 1851-1901. The Victorian Panel Study (VPS). *Historical Social Research* 32, 211-331 (esp. 252-265).

Tan, P.-N, M. Steinbach and V. Kumar (2006). *Introduction to Data Mining.(*Boston: Pearson/Addison Wesley).

8b Main publications of the research team (as far as not included in 8a)

Alter, G., M. Dribe and F. van Poppel (2007). Widowhood, Family Size and Post-Reproductive Mortality: A Comparative Analysis of Three Populations in Nineteenth Century Europe. *Demography*, 44, 4, 785-806.

Bloothooft, G. (1995). Multi-Source Family Reconstruction. *History and Computing* 7, 3, 90-103.

Bruin, J.S. de, T.K. Cocx, W.A. Kosters, J.F.J. Laros and J.N. Kok. (2006). Data Mining Approaches to Criminal Career Analysis, In C.W. Clifton, N. Zhong, J. Liu, B.W. Wah and X. Wu (eds)*Proceedings sixth IEEE international conference on data mining* (Hong Kong: ICDM), 171-177.

Graaf, E.H. de, J.N. Kok and W.A. Kosters (2007). Clustering Improves the Exploration of Graph Mining Results.In C. Boukis, A. Pnevmatikakis and L. Polymenakos (eds.), *Proceedings of the 4th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI07 (New York: Springer)*, 13-20.

Graaf, E.H. de, J. Kazius, J.N. Kok and W.A. Kosters (2007). Visualization and Grouping of Graph Patterns in Molecular Databases. In M. Bramer, F. Coenen and M. Petridis (eds.). *Proceedings of AI-2007, the Twenty-seventh SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence* (Berlin/Heidelberg: Springer), 267-280.

Hoogeboom, H.J., J.F.J. Laros and W.A. Kosters (2008). Selection of DNA Markers *IEEE Transactions on Systems, Man, and Cybernetics* 38, 1, 26-32.

Huijsmans, D.P., N. Sebe (2005). How to Complete Performance Graphs in Content-Based Image Retrieval: Add Generality and Normalize Scope. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 2, 245-251.

Kok, J., K. Mandemakers and H. Wals (2005), City nomads: Changing Residence as a Coping Strategy, Amsterdam, 1890-1940. *Social Science History* 29, 1, 15-44.

Mandemakers, K. (2006). Building Life Course Data Sets from Population Registers by the Historical Sample of the Netherlands (HSN). *History and Computing* 14, 87-108.

Mandemakers, K. and L. Dillon (2004). Best practices with large databases on historical populations. *Historical Methods* 37, 1, 34-38.

9 Requested budget

As instructed by the call, all figures are at 2007 price level and will be adjusted according to norms NWO.

Personal, 2009-2012	**Fte**		
PhD	1		177.495
Programmer	1		211.204
Postdoc IISG	0,5	175.406	
Postdoc Meertens Institute	0,25	80.885	
GENLIAS coordinator	0,3	85.406	
Subtotal IISG, MI, GENLIAS	1,05	341.697	
Paid by NWO	0,8 * 252.676		202.141
Matching by Institutes*		-139.557	
Total personal			**590.840**
Material			
Bench fee PhD			5.000
Hardware **			15.000
Travel budget			30.000
Total material			**50.000**
TOTAL COST			**640.840**

Explanation:

* IISG, Meertens Institute and GENLIAS will match a) 0,25 fte postdoc from the total of 1,05 fte and b) the difference between the low NWO-norms and the actual salaries to be paid.
The matching is calculated as total salary of IISG, Meertens Institute and GENLIAS minus NWO commitment and will be shared in proportion.

IISG	-71.640
Meertens Inst.	-33.035
GENLIAS	-34.882
	-139.557

** A.o. server with 3,20 Ghz, 1.600 Mhz FSB, 2*6 Mb Cache, internal memory: 32 GB including maintenance; a replacement at the end of the term must be made but is not included in the budget.